Published online 1 December 2011 Nucleic Acids Research, 2012, Vol. 40, Database issue D1089–D1092
doi:10.1093/nar/gkr1172

# RecountDB: a database of mapped and count corrected transcribed sequences

**Edward Wijaya[1,2], Martin C. Frith[2], Kiyoshi Asai[1,2] and Paul Horton[2,*]**

[1]Graduate School of Frontier Sciences, University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa 277-8562 and [2]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-41-6, Aomi, Koto-ku, Tokyo 135-0064, Japan

Received August 31, 2011; Accepted November 13, 2011

## ABSTRACT

**The field of gene expression analysis continues to benefit from next-generation sequencing generated data, which enables transcripts to be measured with unmatched accuracy and resolution. But the high-throughput reads from these technologies also contain many errors, which can compromise the ability to accurately detect and quantify rare transcripts. Fortunately, techniques exist to ameliorate the affects of sequencer error. We present RecountDB, a secondary database derived from primary data in NCBI's short read archive. RecountDB holds sequence counts from RNA-seq and 5′ capped transcription start site experiments, corrected and mapped to the relevant genome. Via a searchable and browseable interface users can obtain corrected data in formats useful for transcriptomic analysis. The database is currently populated with 2265 entries from 45 organisms and continuously growing. RecountDB is publicly available at: http://recountdb.cbrc.jp.**

## INTRODUCTION

Sequencing-based transcriptome analysis has been a key area of biological inquiry for over a decade (1). In recent years, next-generation sequencing (NGS) technologies have revolutionized transcriptomics by providing comprehensive, high-resolution measurement of the transcripts in a given biological sample [reviewed in (2,3)]. For example, RNA-seq has been used to uncover new alternative splicing events and discover new genes (4). Furthermore, 5′ capping methods allow exact positional detection of transcription start sites and absolute quantification of transcripts (5).

Despite such benefits, it has been recognized that high-throughput reads from NGS technologies also contain substantial error, for Illumina typically ranging from 0.3% at the beginning of reads to 3.8–25% at the end (6); and that without proper attention this error can cause significant artifacts (7,8).

Unfortunately, simply removing sequences that fail to map well to the genome does not solve this problem, as many genomes (e.g. mammalian) are repetitive enough that erroneous sequences may infortuitously map well to a similar sequence in a different part of the genome. Figure 1 illustrates how the sequencing error combined with a high dynamic range thwarts simple count threshold-based quality control.

To address this problem, many tools have been developed for sequence correction. One class is aimed for genome sequencing (9–12). They assume the true sequences have roughly uniform abundance and identify erroneous sequence by their rareness. This approach is not suitable for transcriptome sequencing; where interesting transcripts (e.g. of transcription factors) often have low copy number. Recent extensions of these methods loosen the assumption of uniformity (13–15) but have not been evaluated for transcriptome analysis.

Fortunately two correction tools have been designed for transcriptome analysis: FreClu (7) and RECOUNT (16). RECOUNT is designed to correct sequence count biases [including for those sequences which should have a zero count (17)] resulting from sequencing error in Solexa/Illumina reads. It uses a probabilistic model to estimate the true expression reads based on their counts and quality scores, without using a reference genome (Figure 1). On a mouse, Solexa/Illumina, 5′ capped transcriptome data set, RECOUNT increased the count of genome mappable sequences by 13.85%, and in some cases the correction qualitatively changed the biological conclusions drawn from the data (16). Although FreClu and RECOUNT use different algorithms, in our evaluation their overall correction was roughly comparable, but RECOUNT uses much less memory and made fewer large sequence count errors when applied to a simulated data set (16).

*To whom correspondence should be addressed. Tel: +81 (3) 3599-8064; Fax: +81 (3) 3599-8081; Email: horton-p@aist.go.jp

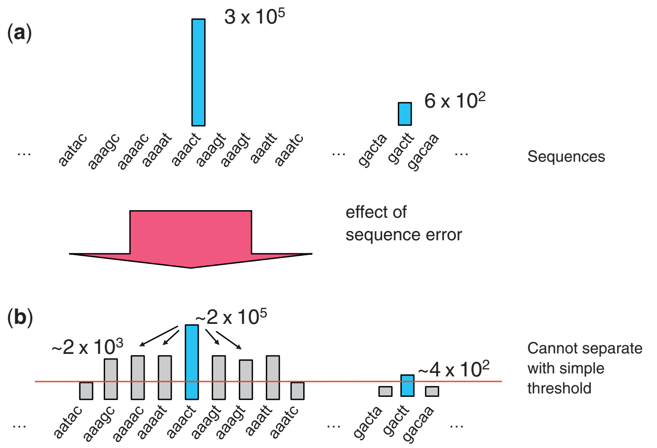

(a)
3 x $10^5$
6 x $10^2$
... aatac aaagc aaaac aaaat aaact aaagt aaagt aaatt aaatc ... gacta gactt gacaa ...
Sequences
effect of sequence error
(b)
~2 x $10^3$
~2 x $10^5$
~4 x $10^2$
... aatac aaagc aaaac aaaat aaact aaagt aaagt aaatt aaatc ... gacta gactt gacaa ...
Cannot separate with simple threshold

**Figure 1.** Effect of sequencing error on the sequence counts. Gray bars represent misread sequences. **(a)** True counts of input sequences; `aaact` and `gactt` are $3 \times 10^5$ and 600, respectively. **(b)** Output of sequencer. Due to sequencer error, misread sequences appear around each input sequence, like crumbs fallen off a cake. Unfortunately, the misreads of highly abundant sequence can have higher count than correct reads of rare sequences. Thus it is in general not possible to separate true and false sequences with a simple threshold. RECOUNT uses a probabilistic model to approximately infer **(a)** from **(b)**.

The NCBI sequence read archive (SRA) plays an important role in preserving experimental data generated from NGS technologies for further studies (18). Typically gene expression levels are obtained from this data by first mapping reads to the reference genome, and then using gene annotation resources such as Aceview (19) or RefSeq (20) to obtain the expression levels of known genes. As discussed above, it is highly desirable to apply sequence count correction before mapping, and convenient for biologists if this can be done for them. For these reasons we developed RecountDB, in which we provide precomputed results of mapping sequences with LAST (21) after sequence correction count with RECOUNT (16), for RNA-Seq and TSS-Seq experiments held in the SRA. RecountDB provides the mapped data in the *de facto* standard formats: PSL (22) and SAM/BAM (23), which can be directly forwarded to visualization and analysis software. Below we describe the database source, content and derivation in detail.

## DATABASE CONTENTS AND FORMATS

RecountDB has been publicly available since June 2010. The current version contains 2265 entries from 45 organisms, with read lengths from 17 bp to 100 bp. Figure 2 shows screen shots of a RecountDB keyword search. Results can be downloaded in three formats: TAB, PSL and SAM/BAM. Below we briefly describe each format.

### TAB format

RecountDB provides a simple format that we call the 'TAB format'. This simple format holds one sequence per line, each line consisting of three tab separated fields: the sequence itself, its observed counts and its corrected count. This format provides minimal information for programmers who want to patch RECOUNT correction into their own genome mapping protocol.

### PSL format

The PSL (Pat Space Layout) format, used by BLAT and other UCSC tools (22), is a tab delimited format consisting of an optional header section and an alignment section. The alignment section consists of 21 fields documented at the UCSC site http://genome.ucsc.edu/FAQ/FAQformat.html#format2L.

RecountDB provides alignments with 14 organisms in UCSC genomes that can be directly uploaded to the UCSC Genome browser (24) as a track. This enables biologists to visualize the transcriptome data together with the extensive annotation on gene structure, regulatory elements etc., available in other tracks.

### SAM/BAM format

RecountDB also provides data in BAM format, a compressed binary version of the SAM (Sequence Alignment/Map) format (23). Many NGS analysis tools work with this format. BAM can be easily converted to SAM format with `SAMtools` (http://sourceforge.net/projects/samtools/files/) (23). This format contains 11 mandatory fields important for alignment including the sequence and quality score information, which is not included in the PSL format.

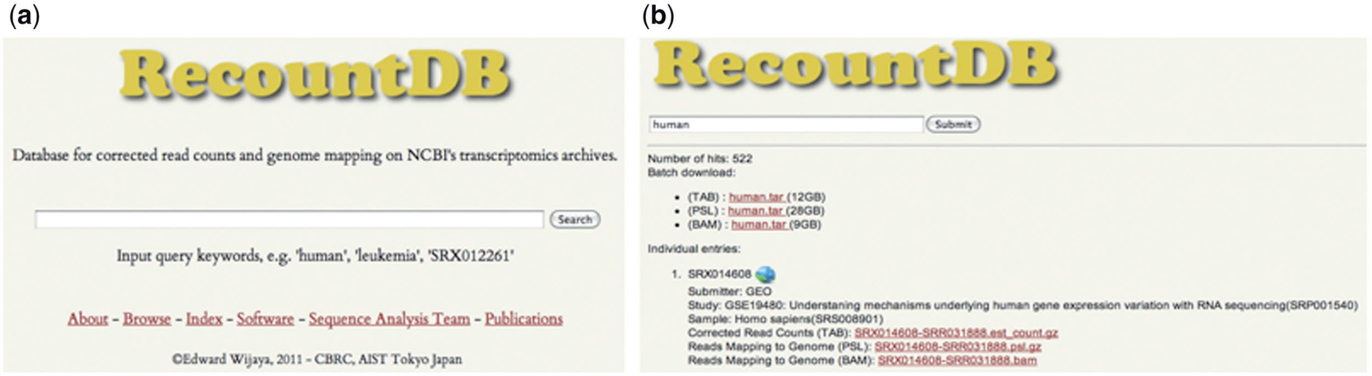
One feature of the SAM format definition is that it allows extension through the addition of extra fields. We take advantage of this by adding the observed and corrected sequence counts as the 14th and 15th fields (Figure 3).

## COMPUTATIONAL PROTOCOL

The NCBI SRA repository holds (NGS) data sets from various experiments. For our purposes we used only data from TSS-seq and RNA-seq experiments. These data sets are required to be submitted in FASTQ format, which encodes the quality of each base in the reads. We make use of the quality scores both for sequence count correction and during the mapping process.

### Error Correction with RECOUNT

RECOUNT (16) adopts the count correction method proposed by Beißbarth *et al.* (25), which is based on a probabilistic model which assumes that the probability of any particular sequence $s$ being misread as some other sequence $r$ is given along with the observed sequence counts. Conceptually, a directed weighted graph $G(V, E)$ is defined in which the vertices $V$ are the possible sequences and the edge weight from sequence $s$ to $r$ represents the probability of misreading $s$ as $r$. As outlined below, the probabilities are automatically derived from quality score information. To keep the computation manageable two approximations are used: (i) sequences with zero observed counts are excluded from consideration, (ii) edges representing very unlikely misread events are omitted, RECOUNT can do this approximately by a hamming distance threshold or more



RecountDB

Database for corrected read counts and genome mapping on NCBI's transcriptomics archives.

Input query keywords, e.g. 'human', 'leukemia', 'SRX012261'

About - Browse - Index - Software - Sequence Analysis Team - Publications

©Edward Wijaya, 2011 - CBRC, AIST Tokyo Japan

RecountDB

human

Number of hits: 522
Batch download:

- (TAB) : human.tar (12GB)
- (PSL) : human.tar (28GB)
- (BAM) : human.tar (9GB)

Individual entries:

1. SRX014608
   Submitter: GEO
   Study: GSE19480: Understaning mechanisms underlying human gene expression variation with RNA sequencing(SRP001540)
   Sample: Homo sapiens(SRS008901)
   Corrected Read Counts (TAB): SRX014608-SRR031888.est_count.gz
   Reads Mapping to Genome (PSL): SRX014608-SRR031888.psl.gz
   Reads Mapping to Genome (BAM): SRX014608-SRR031888.bam

**Figure 2.** RecountDB's search interface. **(a)** The snapshot of the RecountDB entry page. Users can perform searches using keywords such as genome name, or type of study, or NCBI-SRA file ID. A browseable interface can also be accessed through the link in this page. **(b)** A typical RecountDB keyword search result page. Each entry contains basic information such as data submitter, type of study and sample source. The results are presented in three formats: TAB, PSL and SAM/BAM (see main text for explanation). The link in depicted as a globe symbol allows users to reach the NCBI-SRA primary site for the data, where the user can access the original FASTQ file.

```
OC:f:223636   EC:f:264683.909
```

**Figure 3.** Additional fields provided in the RecountDB SAM format data. OC refers to observed count, and EC estimated count (after correction). 'f' refers to type of value (float) and is followed by the values of each type of count.

rigorously by a probability threshold. Currently RecountDB holds results for hamming distance one. Following Beißbarth *et al.* (25), RECOUNT uses a kind of Expectation-Maximization (EM) (26) procedure to infer the set of true counts that locally maximizes the likelihood of the observed reads.

To describe how RECOUNT derives misread probabilities; we define the *consensus sequence* of a read as the sequence called from that read (i.e. the concatenation of the most probable base at each position, using alphabetical order to break ties), and the *reads of a sequence s* as the set of reads whose consensus sequence is *s*. Recall that for any read of *s*, the quality scores give the probability that *s* should have been called as another sequence *r*. Therefore RECOUNT estimates the probability of *r* being misread as *s*, as the harmonic average of that quantity for all reads of *s*. Note that although RECOUNT is capable of utilizing non-uniform base miscall probabilities [for example on Illumina `C` ↔ `T` miscalls are particularly common (6)], the FASTQ data format found in the SRA only gives one quality score for each position. Thus the contents of RecountDB assume miscalled bases are equally likely to be any of the three possible choices.

### Alignment of reads with LAST

*Count corrected mapping*. LAST (http://last.cbrc.jp/) is a local alignment tool designed for fast and accurate genome alignment (21). Using it requires executing two programs: `lastdb` for constructing an index of the genome, and `lastal` for aligning queries using the index. The parameters listed below follow the recommendations of the LAST documentation.

We indexed the reference genome of each relevant species using `lastdb` with option `–m1111110`, which stipulates a spaced seed ignoring mismatches at every seventh position, a pattern suitable for aligning short reads. For `lastal` we use the parameters: `–r6 –q18 –a21 –b9 –d108 –e120`; `–r6` sets the match score to 6, `–q18` the mismatch cost to 18, `–a21` gap existence cost to 21 and `–b9` gap extension cost to 9. The output of these alignments are provided in BAM format.

*Uncorrected mapping*. For users interested in how RECOUNT changed the mapped results, we also provide uncorrected data in PSL format, mapped using LAST quality-score alignment. The parameters are `–Q1 –d108 –e120`. The option `–Q1` tells LAST to expect quality score information in FASTQ-Sanger format. `–d108` sets the minimum score for gapless alignments to 108 and `–e120` sets minimum score for gapped alignments to 120.

*Quality control*. For both types of alignment, we only report alignments with high mapping probabilities. For this purpose we use the LAST script `last-map-probs.py –s150`. With this parameter setting, when mapping 50 bp reads to the human genome we expect a random spurious alignment to be mistakenly reported only once every few thousand reads.

*Format conversion*. The default output of `lastal` is in MAF format. We converted this to PSL and SAM format using the `maf-convert.py` script, from the LAST package.

## CONCLUSIONS

Over the next years the availability of NGS transcriptome data will increase tremendously, covering a growing number of organisms, tissues, cell types and conditions. To obtain reliable conclusions from this data, accurate estimates of transcript abundance are needed. RecountDB can play a valuable role by providing count corrected RNA-Seq and TSS-Seq data in a convenient form.

RecountDB is continuously updated in an automatic fashion. The compact representation of our data (i.e. TAB, PSL and the compressed BAM format) allow us to efficiently store the data. However, the number of amount of data is expected to increase very rapidly. To respond to this, apart from strengthening our hardware infrastructure we also plan to investigate the possibility applying more effective data compression techniques (27).

Finally, as touched upon in the introduction, we note that sequence count correction is an area of intense research. As is true for the other methods, RECOUNT still has room for improvement—for example it assumes the quality scores are approximately accurate. In the future we plan to continue our efforts to improve RECOUNT and updated RecountDB accordingly.

## ACKNOWLEDGEMENTS

We thank Dr Toutai Mituyama for the constructive comments.

## FUNDING

This work is supported by a Japanese Ministry of Education, Sport, Science and Technology (MEXT), Grant-in-Aid for Scientific Research on Innovative Areas (221S002). Funding for open access charge: Grant-in-Aid for Scientific Research on Innovative Areas (221S002).

*Conflict of interest statement*. None declared.

## REFERENCES

1. Velculescu,V.E., Madden,S.L., Zhang,L., Lash,A.E., Yu,J., Rago,C., Lal,A., Wang,C.J., Beaudry,G.A., Ciriello,K.M. *et al.* (1999) Analysis of human transcriptomes. *Nat. Genet.*, **23**, 387–388.
2. Blow,N. (2009) Transcriptomics: the digital generation. *Nature*, **458**, 239–242.
3. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
4. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
5. Carninci,P., Sandelin,A., Lenhard,B., Katayama,S., Shimokawa,K., Ponjavic,J., Semple,C.A., Taylor,M.S., Engstrom,P.G., Frith,M.C. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, **38**, 626–635.
6. Dohm,J.C., Lottaz,C., Borodina,T. and Himmelbauer,H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**, e105.
7. Qu,W., Hashimoto,S. and Morishita,S. (2009) Efficient frequency-based de novo short-read clustering for error trimming in next-generation sequencing. *Genome Res.*, **19**, 1309–1315.
8. Nakamura,K., Oshima,T., Morimoto,T., Ikeda,S., Yoshikawa,H., Shiwa,Y., Ishikawa,S., Linak,M.C., Hirai,A., Takahashi,H. *et al.* (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.*, **39**, e90.
9. Schröder,J., Schröder,H., Puglisi,S.J., Sinha,R. and Schmidt,B. (2009) SHREC: a short-read error correction method. *Bioinformatics*, **25**, 2157–2163.
10. Yang,X., Dorman,K.S. and Aluru,S. (2010) Reptile: representative tiling for short read error correction. *Bioinformatics*, **26**, 2526–2533.
11. Salmela,L. (2010) Correction of sequencing errors in a mixed set of reads. *Bioinformatics*, **26**, 1284–1290.
12. Kelley,D.R., Schatz,M.C. and Salzberg,S.L. (2011) Quake: quality-aware detection and correction of sequencing errors. *Genome Biol.*, **11**, R116.
13. Yang,X., Aluru,S. and Dorman,K.S. (2011) Repeat-aware modeling and correction of short read errors. *BMC Bioinformatics*, **12(Suppl. 1)**, S52.
14. Medvedev,P., Scott,E., Kakaradov,B. and Pevzner,P. (2011) Error correction of high-throughput sequencing datasets with non-uniform coverage. *Bioinformatics*, **27**, i137–i141.
15. Kao,W.C., Chan,A.H. and Song,Y.S. (2011) ECHO: a reference-free short-read error correction algorithm. *Genome Res.*, **21**, 1181–1192.
16. Wijaya,E., Frith,M.C., Suzuki,Y. and Horton,P. (2009) Recount: next generation sequencing error correction tool. *Genome Inform.*, **23**, 189–201.
17. Wijaya,E., Pessiot,J.-F., Frith,M.C., Fujibuchi,W., Asai,K. and Horton,P. (2010) In *Proceedings of the Workshop on data-mining of Next Generation Sequencing Data, BIBM2010*. IEEE Computer Society, Los Alamitos, CA, USA, pp. 561–566.
18. Leinonen,R., Sugawara,H. and Shumway,M. (2011) The sequence read archive. *Nucleic Acids Res.*, **39**, D19–D21.
19. Thierry-Mieg,D. and Thierry-Mieg,J. (2006) AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.*, **7(Suppl. 1)**, S12.
20. Pruitt,K., Tatusova,T. and Maglott,D. (2005) NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D562–D566.
21. Kielbasa,S.M., Wan,R., Sato,K., Horton,P. and Frith,M.C. (2011) Adaptive seeds tame genomic sequence comparison. *Genome Res.*, **21**, 487–493.
22. Kent,W.J. (2002) Blat – the blast-like alignment tool. *Genome Res.*, **12**, 656–664.
23. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
24. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
25. Beißbarth,T., Hyde,L., Smyth,G.K., Job,C., Boon,W.-M., Tan,S.-S., Scott,H.S. and Speed,T.P. (2004) Statistical modeling of sequencing errors in SAGE libraries. *Bioinformatics*, **20(Suppl. 1)**, i31–i39.
26. Dempster,A.P., Laird,N.M. and Rubin,D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.*, **39**, 1–38.
27. Brandon,M.C., Wallace,D.C. and Baldi,P. (2009) Data structures and compression algorithms for genomic sequence data. *Bioinformatics*, **25**, 1731–1738.